P1：

各位评委老师晚上好！我是文华，我们组的项目是，《基于Spark平台的人工智能知识的知识图谱构建》，本次答辩主要由我负责，周余辅助答辩，另一位队友因期末考试无法上线答辩。我们的指导老师是罗月童教授。

P2：

答辩内容分为5个部分，下面开始正式介绍。

P3、P4：

首先是项目背景。Web技术的飞速发展使知识进一步爆炸性地增长，这不可避免地带来了知识冗余的问题。知识图谱作为人工智能领域一项重要的技术分支，具有强大的语义处理能力与开放互联能力，可以帮助用户快速检索、查询所需信息；遗憾的是目前热门的人工智能并没有成熟的知识图谱工具。

P5、P6：

其次是项目简介。当前人工智能方兴未艾、蓬勃发展，交叉性极大、涉及面极广的知识給学习者造成了很大困难。我们选取机器学习、自然语言处理与机器视觉等三个领域作为代表，尝试构建一个人工智能领域的知识图谱。

我们选择Spark作为数据处理的工具，适用Jiagu模型进行知识关系抽取，CSDN与博客园是主要的元数据来源。

P7、P8：

接下来介绍项目的完成情况。

首先是技术部分。

第一：数据获取。我们使用Python编写爬虫程序从上面提到的CSDN和博客园爬取网页文档，为了提高爬取效率，我们使用了主从分布式爬虫结构：将一台阿里云服务器作为master服务器，负责维护数据列表与作业分发；由三台本地的笔记本电脑组成slave服务器组，进行具体的爬取任务，slave服务器间不进行通信。

P9：

第二：数据预处理。使用Spark与hive对存储在数据库中数据进行快速地进行抽取、加载与转换。使用PPT所示的几种API完成数据的存储、去重、标签过滤与敏感词去除。

P10：

第三：知识抽取。使用Jiagu模型对文本进行关系抽取，为了得到适用于本项目的模型，我们调整了相关参数对标记文本进行训练，实验结果用held-out方法进行评估，各位评委可以在支撑材料中查阅有关实验更加详细的介绍。

P11：

第四：图谱可视化。使用amCharts 4作为可视化工具，通过PHP处理后台三元组数据，将满足要求的json文本返回到前端，由amCharts 4完成可视化。

下面是团队协作与老师指导部分。

P12：

这是我们大创的分工与预期时间分配情况。

P13：

这是我们的经费使用情况，显然在保障项目顺利完成的基础上，项目经费并未超支。购买的内存条已经得到了充分的利用，因疫情而无法返校，不能将显卡装在实验室的机子上而为我们项目的数据训练提供帮助。

P14：

这是在整个项目实施过程中我们与指导老师的互动情况。

P15、P16：

接下来介绍项目创新。

我们构建的知识图谱工具采用力导向图将三元组表示的数据展开为各个知识节点，并添加了统计信息，同时支持用户选择多样的主题背景。

P17-P21：

这些是图谱可视化的具体形式。

P22：

最后，我们都很感谢参加这次比赛的机会，让我们各方面的能力都得到了锻炼与提高。

P23：

请各位评委老师批评指正！